OPERATIONS RESEARCH
Vol. 58, No. 4, Part 1 of 2, July–August 2010, pp. 889–901
ISSN 0030-364X | EISSN 1526-5463 | 10 | 5804 | 0889

informs

DOI 10.1287/opre.1090.0773

On Choosing Parameters in Retrospective-Approximation Algorithms for Stochastic Root Finding and Simulation Optimization

Raghu Pasupathy
Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia 24061, pasupath@vt.edu

The stochastic root-finding problem is that of finding a zero of a vector-valued function known only through a stochastic simulation. The simulation-optimization problem is that of locating a real-valued function's minimum, again with only a stochastic simulation that generates function estimates. Retrospective approximation (RA) is a sample-path technique for solving such problems, where the solution to the underlying problem is approached via solutions to a sequence of approximate deterministic problems, each of which is generated using a specified sample size, and solved to a specified error tolerance. Our primary focus, in this paper, is providing guidance on choosing the sequence of sample sizes and error tolerances in RA algorithms. We first present an overview of the conditions that guarantee the correct convergence of RA's iterates. Then we characterize a class of error-tolerance and sample-size sequences that are superior to others in a certain precisely defined sense. We also identify and recommend members of this class and provide a numerical example illustrating the key results.

Subject classifications: simulation: efficiency, design of experiments; programming: stochastic.
Area of review: Simulation.
History: Received March 2007; revisions received November 2007, August 2008, September 2008, October 2008; accepted November 2008. Published online in *Articles in Advance* March 24, 2010.

1. Introduction and Motivation

The stochastic root-finding problem (SRFP) and the simulation optimization problem (SOP) are simulation-based stochastic analogues of the well-researched root-finding and optimization problems, respectively. These problems have recently generated a tremendous amount of attention amongst researchers and practitioners, primarily owing to their generality. Because the functions involved in these formulations are specified implicitly through a stochastic simulation, virtually any level of complexity is afforded. Various flavors of SRFP and SOP have thus found application in an enormous range of large-scale, real-world contexts such as vehicular transportation networks, quality control, telecommunication systems, and health care. See Andradóttir (2006), Spall (2003), Fu (2002), Barton and Meckesheimer (2006), Chen and Schmeiser (2001), and Ólafsson (2006) for entry points into this literature and overviews on the subject.

For solving SOPs and SRFPs, *sample-average approximation* (SAA), amongst a few other popular classes of algorithms (see §3), has found particular expediency amongst researchers and practitioners. In its most basic form, SAA involves generating a sample-path problem using an appropriately chosen sample size, and then solving it to desired precision using a chosen numerical procedure. This step is usually followed by an analysis of the solution estimator, using the now well-established large-sample (Robinson 1996, Shapiro 2004, Atlason et al. 2004) and small-sample properties (Mak et al. 1999), and theories on assessing solution quality (Bayraksan and Morton 2007). Due to its simplicity, and the ability to incorporate powerful and mature tools from the deterministic context, SAA has been widely applied. Some examples of SAA application by experts include optimal release times in production flow lines (Homem-de-Mello et al. 1999), call-center staffing (Atlason et al. 2004, 2008), tandem production lines and stochastic PERT (Plambeck et al. 1996), transshipment problems (Herer et al. 2006), design for health care (Prakash et al. 2008), and vehicle-routing problems (Verweij et al. 2003).

Our specific focus in this paper is a recent refinement of the SAA class of methods called, variously, retrospective approximation (RA) (Chen and Schmeiser 2001, Pasupathy and Schmeiser 2009) and variable sample size (Homem-de-Mello 2003) methods. Whereas a generic SAA method generates a *single* sample-path problem with a large enough sample size, and then solves it to a prescribed error tolerance, the said SAA refinements generate a *sequence* of sample-path problems with progressively increasing sample sizes, and then solve these to progressively decreasing error tolerances. This elaborate structure in the SAA refinements is explicitly constructed to gain overall efficiency. The early iterations are efficient, in principle, because the small sample sizes ensure that not much computing effort

is expended in generating sample-path problems. The later iterations are efficient, again in principle, because the starting solution for the sample-path problem is probably close to the true solution, and not much effort is expended in solving sample-path problems. The solving of the individual sample-path problems, as in generic SAA, is accomplished by choosing any numerical procedure from amongst the powerful host of deterministic root-finding/optimization techniques that are currently available.

The general RA structure is indeed attractive from both the efficiency and the implementability perspectives, and the resulting RA estimators inherit much of the well-studied large-sample properties of generic SAA estimators. Little is currently known, however, about how exactly to trade off parameters, i.e., sample sizes and error tolerances, within such frameworks. Although these frameworks seem to serve as viable implementation refinements to SAA, the stipulations dictated by convergence allow an enormous number of possible choices for sample-size and error-tolerance sequences, some of which are conceivably much inferior to others from an algorithm efficiency standpoint. It seems intuitively clear that the sample-size and error-tolerance sequences should be chosen in balance—increasing sample sizes too fast compared to the decrease rate of error tolerances will lead to residual bias from the undersolving of sample-path problems; increasing sample sizes too slowly compared to the decreasing rate of error tolerances will lead to wasted computational effort resulting from the oversolving of the sample-path problems. This trade-off resulting from the choice of sample sizes and error tolerances motivates our central questions:

—Does there exist a balanced choice of sample sizes and error tolerances in SAA refinements such as RA, where the terms "sample size" and "error tolerance" refer to some generic measures of problem-generation effort and solution quality, respectively? Furthermore, can this balance be characterized rigorously?

—Can the characterization of a balanced choice be used to provide guidance in choosing particular sample-size and error-tolerance sequences automatically across iterations within SAA refinements such as RA?

As we shall see, the answer to both of the above questions is in the affirmative. Why is parameter choice such an important question? In broad terms, parameter choice is intimately linked to algorithm efficiency. Frequently, SOPs and SRFPs arise in contexts where simulations are time consuming, taking anywhere from a few seconds to a few hours for each run. In such situations, algorithm parameter choices can have a dramatic effect on solution quality as a function of time or computing effort. Moreover, the researcher/practitioner often does not have the flexibility to tweak algorithm parameters so as to identify good parameters. This may be because of the lack of expertise, or simply because solutions need to be obtained in a rapid and automatic fashion. Proper guidance on the choice of parameters is thus imperative to ensure that resources are utilized in an efficient fashion within RA-type algorithms. Considering applicability to a wide audience, such guidance is especially useful if identified within a generic framework such as we consider in this paper. Optimal parameter choice is not an RA-specific issue. It has been recognized as an important question in other algorithm classes as well, e.g., stochastic approximation algorithms (Spall 2003, 2006).

1.1. Contributions

That SAA is currently amongst the attractive methods of solving SRFPs and SOPs is undeniable. The enormous amount of recent theoretical development (Bayraksan and Morton 2007, 2010; Chen and Schmeiser 2001; Higle and Sen 1991; Homem-de-Mello 2003; Kleywegt et al. 2001; Mak et al. 1999; Plambeck et al. 1996; Polak and Royset 2008; Shapiro 2000, 1991, 2004), and its expediency in a wide range of actual applications (Homem-de-Mello et al. 1999; Atlason et al. 2004, 2008; Plambeck et al. 1996; Herer et al. 2006; Prakash et al. 2008; Verweij et al. 2003), stand in evidence. Considering this broad appeal of SAA, we address key questions relating to how parameters, specifically sample sizes and error tolerances, should be chosen within implementable refinements of SAA. Parameter choice is crucial in that it plays an important role in deciding the efficiency of the solutions resulting from SAA algorithm execution.

The following are specific contributions of this paper.

1. We rigorously characterize an *optimal class* of sample sizes and error tolerances for use within sample-path methods. The characterization of this *optimal class* is in the form of three limiting conditions involving the parameter sequences, and the convergence rate of the numerical procedure used to solve the sample-path problem. We show that the characterization is tight in the sense that optimality, as defined, is achieved if and only if the chosen parameters fall within the characterized class.

2. With a view toward actual implementation, we identify commonly considered classes of sample-size and error-tolerance sequences that lie inside and outside the characterized optimal class. For sublinear, linear, and polynomially converging numerical procedures used within sample-path problems, we provide guidance on the rate at which sample sizes should be increasing if one is looking to choose parameters that belong to the optimal class.

3. For completeness and ease of exposition, we clarify the conditions under which RA algorithms converge to the correct solution with probability one (wp1). In addition, we prove a certain central limit theorem for RA estimators on SRFPs. Similar results already exist for SOPs.

1.2. Organization

The remainder of the paper is organized as follows. In §2 we provide the problem statements for SRFPs and SOPs, followed by some notation and terminology used in this paper. In §3 we present a brief literature review on the

existing methods to solve SRFPs and SOPs. Convergence results on RA algorithms appear in §4, followed by the main results in §5. Some of these results, especially those appearing in §4, have been stated in abbreviated form or without proofs in Pasupathy (2006). In §6, we discuss our choice of efficiency measure and termination criteria. Section 7 includes an illustration of the results obtained in the paper through a numerical example. We provide concluding remarks in §8.

2. Problem Statements and Notation

The following is a list of key notation and definitions adopted in the paper: (i) x^* denotes a *true solution* to the SRFP or SOP; (ii) X_k^* denotes a *true solution* to the kth sample-path problem; (iii) X_k denotes the kth *retrospective solution*, i.e., the estimated solution to the kth sample-path problem; (iv) $X_n \xrightarrow{p} X$ means that the sequence of random variables $\{X_n\}$ converges to the random variable X in probability; (v) $X_n \to X$ wp1 means that the sequence of random variables $\{X_n\}$ converges to the random variable X with probability one; (vi) $X_n \xrightarrow{d} X$ means that the sequence of random variables $\{X_n\}$ converges to the random variable X in distribution; (vii) $\{m_k\} \uparrow \infty$ means that the sequence $\{m_k\}$ is an increasing sequence going to infinity; (viii) the sequence $\{m_k\} \to \infty$ exhibits *sublinear growth* if $\limsup_{k\to\infty} m_k/m_{k-1} \leqslant 1$, *linear growth* if $1 < \limsup_{k\to\infty} m_k/m_{k-1} < \infty$, *polynomial growth* if $0 < \limsup_{k\to\infty} m_k/(m_{k-1})^p < \infty$ for some $p > 1$, and *exponential growth* if $\limsup_{k\to\infty} m_k/(m_{k-1})^p = \infty$ for all $p > 1$; (ix) $\lceil x \rceil$ denotes the smallest integer greater than or equal to $x \in \mathbb{R}$; (x) $\mathrm{dist}(x, \Omega) = \inf\{\|x - y\|\colon\ y \in \Omega\}$ denotes distance between a point $x \in \mathbb{R}$ and a set Ω; (xi) $B(x, r)$ denotes a ball of radius r centered on x.

Formally, the SRFP is stated as follows.

Given: A simulation capable of generating, for any $x \in \mathfrak{D} \subset \mathbb{R}^q$, an estimator $\bar{Y}_m(x)$ of the function $g\colon \mathfrak{D} \to \mathbb{R}^q$ such that $\bar{Y}_m(x) \xrightarrow{d} g(x)$ as $m \to \infty$, for all $x \in \mathfrak{D}$.

Find: A zero $x^* \in \mathfrak{D}$ of g, i.e., find x^* such that $g(x^*) = 0$, assuming that one such exists.

Similarly, the version of SOP we will use in this paper is as follows.

Given: A simulation capable of generating, for any $x \in \mathfrak{D} \subset \mathbb{R}^q$, an estimator $\bar{Y}_m(x)$ of the function $g\colon \mathfrak{D} \to \mathbb{R}$ such that $\bar{Y}_m(x) \xrightarrow{d} g(x)$ as $m \to \infty$, for all $x \in \mathfrak{D}$.

Find: A local minimizer $x^* \in \mathfrak{D}$ of g, i.e., find x^* having a neighborhood $V(x^*)$ such that every $x \in V(x^*)$ satisfies $g(x) \geqslant g(x^*)$, assuming that one such x^* exists.

As stated, the SRFP and SOP make no assumptions about the nature of $\bar{Y}_m(x)$ except that $\bar{Y}_m(x) \xrightarrow{d} g(x)$ as $m \to \infty$. Also, the feasible set $\mathfrak{D}$ is assumed to be known in the sense that the functions involved in the specification of $\mathfrak{D}$ are observed without error. Various slightly differing flavors of the SOP have appeared in the literature. See, for example, Nemirovski and Shapiro (2004).

3. Abbreviated Literature Review

In this section, we present a brief overview of the important works related to solving SRFPs and SOPs. We limit ourselves to a broad categorization, providing references only to summary articles. Sample-average approximation, being the topic of this paper, is discussed in greater detail.

Current methods to solve SOPs fall into six broad categories: (i) metamodels, (ii) metaheuristics, (iii) comparison methods, (iv) random-search methods, (v) stochastic approximation (SA), and (vi) sample-average approximation (SAA). This categorization is not mutually exclusive, but methods in each category share some set of distinguishing features. For instance, metamodels (category (i)) estimate a functional relationship between the objective function and the variables in the search space, and then employ deterministic optimization techniques on the estimated objective function. See Barton and Meckesheimer (2006) for detailed ideas. The methods in category (ii) are based on metaheuristics such as tabu search, evolutionary algorithms, simulated annealing, and nested partitions (Ólafsson 2006). Comparison methods (category (iii)), a mature class of methods predominantly based on statistical procedures, have been specially developed to solve SOPs where the set of feasible solutions is discrete and small. See Kim and Nelson (2006) for a broad overview. Random-search methods (category (iv)) include algorithms that use some intelligent sampling procedure to traverse the search space, followed by an update mechanism to report the current best solution. See Andradóttir (2006) for an overview. SA methods (category (v)) are based on a recursion introduced in Robbins and Monro (1951), and have seen enormous development. Several books (Kushner and Clark 1978, Kushner and Yin 2003, Spall 2003) are available on the subject. Multiscale SA (Bhatnagar and Borkar 1998, Bhatnagar et al. 2001, Tadic and Meyn 2003), an interesting subclass of SA, improves on generic SA using paired recursions for parameter updating and system evolution.

SAA (category (vi)) is the category of most interest in this paper. SAA techniques seem to have first appeared in Shapiro (1991) and Healy and Schruben (1991) and were later used by several other authors (Rubinstein and Shapiro 1993, Plambeck et al. 1996, Atlason et al. 2004) in various contexts. The idea of SAA is easily stated. Obtain an estimator for the true solution x^* of the SOP (or SRFP) by solving an approximate problem S. The approximate problem S substitutes the unknown function g in the original problem by its sample-path approximation $\bar{y}_m(x; \underline{\omega})$ generated with sample size m, and the vector of random numbers $\underline{\omega}$. SAA has a well-developed small-sample (Mak et al. 1999) and large-sample (Robinson 1996, Shapiro 2004, Kleywegt et al. 2001, Atlason et al. 2004) theory. Recent work includes methods to assess solution quality (Bayraksan and Morton 2007).

Retrospective approximation (RA) (Chen and Schmeiser 2001, Pasupathy and Schmeiser 2009) is a variant of SAA where instead of generating and solving a *single* sample-path problem S, a *sequence* of sample-path problems $\{S_k\}$ are generated with sample sizes $\{m_k\} \to \infty$, and solved to error tolerances $\{\epsilon_k\} \to 0$. The estimated solution X_{k-1} of the sample-path problem S_{k-1}, also called the $(k-1)$th *retrospective solution*, is used as the initial solution to the subsequent sample-path problem S_k. The philosophy behind the RA structure is as follows: During the early iterations, use "small" sample-sizes m_k and "large" error-tolerances ϵ_k in solving the sample-path problem S_k; in later iterations, as the retrospective solution X_k tends closer to the true solution x^*, use "larger" sample sizes and "smaller" error tolerances. In implementing RA algorithms, three parameters need to be chosen: (i) a numerical procedure for solving the sample-path problem S_k; (ii) a sequence of sample sizes $\{m_k\}$ to be used for generating the sample-path functions $\{\bar{y}_{m_k}(x; \underline{\omega}_k)\}$; and (iii) a sequence of error tolerances $\{\epsilon_k\}$ to guide termination of iterations. The subject of this paper is the rigorous choice of parameters (i), (ii), and (iii), with an emphasis on (ii) and (iii).

Three recent papers are directly relevant to what we discuss in this paper. In Homem-de-Mello (2003), through a setup that is very similar to that considered in this paper, stipulations on sample-size growth rates are identified so as to ensure the consistency of the objective function estimator. The results derived in this paper thus complement those in Homem-de-Mello (2003). In Polak and Royset (2008), through a framework that is slightly less generic than that considered in this paper, an efficient schedule of sample sizes and stages is identified through the solution of an auxiliary optimization problem, and under the assumption that the procedure used to solve the sample-path problems is linearly convergent. Most recently, in Bayraksan and Morton (2010), a sequential sampling procedure is suggested for use within SAA-type algorithms. The primary contribution of Bayraksan and Morton (2010) is the identifying of a schedule of sample sizes, along with the set of conditions, sufficient to provide probabilistic guarantees on the optimality gap of a candidate solution obtained through SAA. Differences between the current work and Bayraksan and Morton (2010) are worth noting—whereas the broad objective in Bayraksan and Morton (2010) is obtaining an SAA solution that is of a guaranteed prespecified quality (in a probabilistic sense), ours is ensuring that the sequence of obtained SAA solutions converges optimally. For the same reason, our results are intimately linked to the convergence rate of the numerical procedure used to solve the sample-path problems within SAA, unlike in Bayraksan and Morton (2010), where the optimality gap is estimated. The "controllables" in both the current paper and Bayraksan and Morton (2010) are the sample sizes used in generating the sample-path problems.

4. Conditions for Guaranteed Convergence

In this section, we state sufficient conditions to ensure that the sequence $\{X_k\}$ of retrospective solutions in RA algorithms converges to the true solution x^* wp1. Separate treatments are provided for the SRFP and the SOP contexts. The results presented in this section are a simple consequence of some well-known results in the literature.

We start with Theorem 1, which appears as Theorem 5.1 in Shapiro (2000). Theorem 1 asserts that under mild conditions, a sequence $\{X_k^*\}$ of *global minimizers* of the sample-path functions $\{\bar{y}_{m_k}(x; \underline{\omega}_k)\}$ converges in distance to the set π^* of *global minimizers* of the limiting function g wp1.

THEOREM 1. *Assume that* (i) *the set* $\mathfrak{D} \subset \mathbb{R}^q$ *is compact*; (ii) *the function* $g\colon \mathbb{R}^q \to \mathbb{R}$ *is continuous*; *and* (iii) *the functional sequence* $\{\bar{y}_{m_k}(x; \underline{\omega}_k)\}$ *converges to* g *uniformly wp*1. *Then, if* X_k^* *is a global minimizer of* $\bar{y}_{m_k}(x; \underline{\omega}_k)$, *and* π^* *is the nonempty set of global minimizers of* g, *the distance* $dist(X_k^*, \pi^*) \to 0$ *wp*1.

Theorem 1 is more useful in proving convergence of RA iterates for the SRFP context than for the SOP context. This is because although we seek a local minimizer of the limit function g in SOPs, Theorem 1 is global in nature. In other words, Theorem 1 talks about the convergence of *global minimizers* of the sample-path functions to the set of *global minimizers* of the limit function g.

4.1. SRFP Context

To establish convergence of the retrospective solutions $\{X_k\}$ to the true solution x^* in the SRFP context, we first convert the original SRFP into an equivalent SOP where every local minimizer is also a global minimizer, and then apply Theorem 1. This is demonstrated in Theorem 2, the main convergence result for RA iterates in the context of SRFPs.

THEOREM 2 (SRFP). *Assume that* (i) *the set* $\mathfrak{D} \subset \mathbb{R}^q$ *is compact*; (ii) *the function* $g\colon \mathbb{R}^q \to \mathbb{R}^q$ *is continuous*; *and* (iii) *the functional sequence* $\{\bar{y}_{m_k}(x; \underline{\omega}_k)\}$ *converges to* g *uniformly wp*1. *Let the positive-valued sequence* $\{\epsilon_k\} \to 0$, *and* X_k *satisfy* $\|X_k - X_k^*\| \leqslant \epsilon_k$, *where* X_k^* *is a zero of* $\bar{y}_{m_k}(x; \underline{\omega}_k)$. *Then,* $dist(X_k, \pi^*) \to 0$ *wp*1, *where* π^* *is the set of zeros of the function* g.

PROOF. See appendix. □

4.2. SOP Context

Theorem 1 is a statement about the behavior of a sequence of global minimizers of sample-path functions. In studying SOPs, Theorem 1 is thus useful only if the numerical procedure used to solve the sample-path problems returns a global minimizer. Our interest, however, is in analyzing RA algorithms where sample-path problems are solved for a local minimizer, and that too with finite stopping. For such situations, Theorem 1 is not as useful because it says nothing about the behavior of any particular sequence of

local minimizers. For instance, consider $g(x) = |x|$, and $\bar{y}_{m_k}(x; \underline{\omega}_k) = g(x) + f_k(x)I_{[1-1/k,\,1+1/k]}$, where f_k is any continuous function with a unique minimum at $x = 1$, $f_k(x) \leqslant 0$ for all $x \in [1 - 1/k, 1 + 1/k]$, $f_k(1 - 1/k) = f_k(1 + 1/k) = 0$, and $\lim_{k\to\infty} f_k(1) = 0$ (e.g., $f_k(x) = (1-x)^2 - (1/k)^2$). Then, $\bar{y}_{m_k}(x; \underline{\omega}_k)$ has the same two local minima for each k and differs from $g(x)$ only in the interval $(1 - 1/k, 1 + 1/k)$. Also, the sequence $\{\bar{y}_{m_k}(x; \underline{\omega}_k)\}$ converges to $g(x)$ uniformly, and the set of global minimizers of the function $\bar{y}_{m_k}(x; \underline{\omega}_k)$ converges to the global minimizer of $g(x)$ as $k \to \infty$. Notice, however, that the sequence $\{x_k = 1\}$ of local minimizers of $\{\bar{y}_{m_k}(x; \underline{\omega}_k)\}$ does not converge to a local minimizer of the limit function g. This example implies, as Homem-de-Mello (2003) notes, that using locally convergent algorithms to solve sample-path problems may "trap" SAA solutions at a local minimizer. In other words, even in the hypothetical scenario where a local minimizer is solved to infinite precision during each iteration, there is no guarantee that the obtained sequence of local minimizers will converge to a true solution x^*.

The following is a technical assumption devised by Bastin et al. (2006) to exclude possibilities of the type illustrated in the previous example. We call it the *rigidity* assumption to reflect the fact that it excludes such examples by stipulating that local minimizers of sample-path functions that persist across iterations, do so over some arbitrarily small ball whose size remains *rigid*, i.e., whose radius remains larger than some threshold.

ASSUMPTION 1. *Let $\mathfrak{L}_k(\underline{\omega}_k)$ be the set of local minimizers of $\bar{y}_{m_k}(x; \underline{\omega}_k)$, and let $\mathfrak{L}^*$ be the set of cluster points of all sequences $\{X_k^*\}$, $X_k^* \in \mathfrak{L}_k(\underline{\omega}_k)$. If $l^* \in \mathfrak{L}^*$, then there exists a subsequence $\{X_{k_j}^*\} \subseteq \{X_k^*\}$ converging to l^*, and constants $\delta > 0$, $t > 0$, such that $X_{k_j}^*$ is a minimizer of $\bar{y}_{m_k}(x; \underline{\omega}_k)$ in $B(X_{k_j}^*, \delta)$ for all $j > t$.*

This assumption has the same implications as in Bastin et al. (2006)—it prevents the occurrence of kinks in the sample-path functions that vanish with increasing sample size. This is a reasonable stipulation because such examples where local minima appear artificially in the sample-path problems and then disappear as the sample size is increased seem infrequent in practice. Theorem 4.1 in Bastin et al. (2006) then directly applies to any sequence of local minimizers $\{X_k^*\}$ of the sample-path functions $\{\bar{y}_{m_k}(x; \underline{\omega}_k)\}$. Specifically, it guarantees that when Assumption 1 holds, $\{X_k^*\}$ converges to some local minimizer x^* of the limit function g wp1. Now recall that the retrospective solution X_k identified during the kth iteration in RA is at most ϵ_k away from X_k^*. Because $\{\epsilon_k\} \to 0$, the convergence of $\{X_k\}$ to some local minimizer x^* of g becomes a trivial extension of Theorem 4.1 in Bastin et al. (2006). We formally state this next without a proof.

THEOREM 3 (SOP). *Let the conditions in Theorem* 1 *and Assumption* 1 *hold. Let the positive-valued sequence $\{\epsilon_k\} \to 0$, and X_k satisfy $\|X_k - X_k^*\| \leqslant \epsilon_k$, where X_k^* is a local minimizer of $\bar{y}_{m_k}(x; \underline{\omega}_k)$. Then, dist$(X_k, \pi^*) \to 0$ wp*1*, where π^* is the set of local minimizers of the function g.*

5. Choosing Sample-Size and Error-Tolerance Sequences

Having discussed convergence of RA iterates in the previous section, our objective in this section is more interesting—provide rigorous guidance in choosing RA algorithm parameters, specifically the error-tolerance sequence $\{\epsilon_k\}$ and the sample-size sequence $\{m_k\}$. We wish to automatically choose the sequences $\{\epsilon_k\}$ and $\{m_k\}$ in a fashion that ensures that the retrospective solutions $\{X_k\}$ converge to the true solution x^* in some reasonably and precisely defined optimal sense.

Before we present the main ideas on choosing parameter sequences, we present a type of central limit theorem (CLT) on the sample-path solution X_k^* in the context of SRFPs, essential for later exposition. A corresponding result for the context of SOPs appears in Shapiro (2000, Theorems 5.2, 5.3, 5.4). We follow the notation established in the previous section—$\bar{y}_{m_k}(x; \underline{\omega}_k)$ is the kth sample-path function generated with sample size m_k, the sample-path solution X_k^* is a zero of $\bar{y}_{m_k}(x; \underline{\omega}_k)$ in the context of SRFPs and some local minimizer of $\bar{y}_{m_k}(x; \underline{\omega}_k)$ in the context of SOPs, and the retrospective solution X_k is obtained by solving the kth sample-path problem to within the chosen tolerance ϵ_k. A proof is provided in the appendix.

THEOREM 4 (SRFP). *Let the conditions of Theorem* 2 *hold. Furthermore, assume* (i) *X_k^* and x^* are the unique zeros of the functions $\bar{y}_{m_k}(x; \underline{\omega}_k)$ and $g(x)$, respectively*; (ii) *the functions $\bar{y}_{m_k}(x; \underline{\omega}_k)$ and $g(x)$ have nonsingular derivatives $\nabla\bar{y}_{m_k}(x; \underline{\omega}_k)$, $\nabla g(x)$ in some neighborhood around x^**; (iii) *the sequence $\{\nabla\bar{y}_{m_k}(x; \underline{\omega}_k)\}$ converges uniformly (elementwise) to $\nabla g(x)$ in some neighborhood around x^* wp*1; *and* (iv) *a central limit theorem holds for $\bar{y}_{m_k}(x; \underline{\omega}_k)$, i.e., $\sqrt{m_k}(\bar{y}_{m_k}(x; \underline{\omega}_k) - g(x)) \xrightarrow{d} N(0, \Sigma)$, where $N(0, \Sigma)$ is the Gaussian random variable with mean zero and covariance Σ. Then,*

$$\sqrt{m_k}(X_k^* - x^*) \xrightarrow{d} N(0, \nabla g(x^*)^{-1}\Sigma(\nabla g(x^*)^{-1})^T).$$

The assumption (iii) on the sequence of derivatives $\{\nabla\bar{y}_{m_k}(x; \underline{\omega}_k)\}$ is important and ensures that the sample-path functions smoothly approximate the true function. Under what conditions can this assumption be expected to hold? The most general answer to this question is provided by standard results in most real-analysis textbooks. For example, Theorem 7.17 in Rudin (1976, p. 152) guarantees that the uniform convergence of $\{\nabla\bar{y}_{m_k}(x; \underline{\omega}_k)\}$ is sufficient to ensure that $\{\nabla\bar{y}_{m_k}(x; \underline{\omega}_k)\}$ converges to $\nabla g(x)$. Similarly, when $g(x)$ is an expectation, the uniform integrability of the finite differences formed from $\{\bar{y}_{m_k}(x; \underline{\omega}_k)\}$, along with the finiteness of $g(x)$, provide

necessary and sufficient conditions for $\{\nabla \bar{y}_{m_k}(x; \underline{\omega}_k)\} \to \nabla g(x)$ to hold (Glasserman 1991, p. 14). The most popular method to guarantee $\{\nabla \bar{y}_{m_k}(x; \underline{\omega}_k)\} \to \nabla g(x)$ when $g(x)$ is an expectation is through Lebesgue's dominated convergence theorem (Rudin 1976, p. 321) in combination with the generalized mean value theorem (Glasserman 1991, p. 15).

These results, although providing conditions in all generality, still do not leave us with complete intuition on whether $\{\nabla \bar{y}_{m_k}(x; \underline{\omega}_k)\} \to \nabla g(x)$ will be satisfied for a given system. This question of when $\{\nabla \bar{y}_{m_k}(x; \underline{\omega}_k)\} \to \nabla g(x)$ is addressed in great detail in Glasserman (1991). The key factor turns out to be the continuity of the performance measure $\bar{y}_{m_k}(x; \underline{\omega}_k)$. Although neither necessary nor sufficient to ensure $\{\nabla \bar{y}_{m_k}(x; \underline{\omega}_k)\} \to \nabla g(x)$, the continuity of $\bar{y}_{m_k}(x; \underline{\omega}_k)$ ensures that only a few other mild assumptions are needed to deem the assumption valid. For example, let $g(x) = \mathrm{E}[\bar{y}_{m_k}(x; \underline{\omega}_k)]$. Then, if $\bar{y}_{m_k}(x; \underline{\omega}_k)$ is continuously differentiable on a compact set C and $\mathrm{E}[\sup_{x \in C} \nabla \bar{y}_{m_k}(x; \underline{\omega}_k)] < \infty$, the interchange holds, i.e., $\mathrm{E}[\nabla \bar{y}_{m_k}(x; \underline{\omega}_k)] = \nabla g(x)$ on C. Furthermore, in generalized semi-Markov processes (GSMP)—a general model for discrete-event systems seen in practice—the stipulation of continuity in $\bar{y}_{m_k}(x; \underline{\omega}_k)$ can be checked through a commuting condition on the GSMP representation of the system in consideration (Glasserman 1991, Chapter 3).

Various other specific contexts have routinely used $\{\nabla \bar{y}_{m_k}(x; \underline{\omega}_k)\} \to \nabla g(x)$, particularly when $g(x)$ is an expectation. For example, the said exchange has been validated in great detail in fluid-flow models (Wardi et al. 2002). In stochastic linear programs (Higle and Sen 1991), the structure of the problem automatically produces continuous sample paths, thereby naturally allowing exchanges of this type.

The other important assumption (i) in Theorem 4—the functions $\bar{y}_{m_k}(x; \underline{\omega}_k)$ and $g(x)$ have unique zeros—is admittedly stringent. Two remarks are relevant regarding this assumption. First, it has been our experience that in SOP and SRFP contexts, sample paths almost invariably mimic the structural properties of the limiting function. (A notable exception is an empirical cumulative distribution function.) This is indeed interesting and lends some credibility to assuming similar properties on $\bar{y}_{m_k}(x; \underline{\omega}_k)$ and $g(x)$. Second, virtually all Newton-based iterations to find zeros of a function seem to need an assumption on the nearness of the starting point (to a zero) in order to prove convergence. Our assumption of a unique zero is comparable, and can be relaxed through similar assumptions on the closeness of the starting solution when solving a sample-path problem.

5.1. Relation Between Sample-Size and Error-Tolerance Sequences

Our message is that for optimal convergence, choose $\{\epsilon_k\}$ and $\{m_k\}$ in balance so that they converge to their respective limits, zero and infinity, at similar rates. We make this claim precise through Theorem 5, where we prove that those sample-size and error-tolerance sequences that satisfy conditions C.1, C.2, and C.3, to be stated, are superior to others in a certain precise sense.

Before we state conditions C.1, C.2, and C.3, we remind the reader of the notions of sublinear, linear, and superlinear convergence. Let the deterministic sequence $\{z_k\} \to z^*$ with $z_k \neq z^*$ for all k. Then the quotient convergence factors, or Q-factors, of the sequence $\{z_k\}$ are $Q_p = \limsup_{k\to\infty} \|z_{k+1} - z^*\| / \|z_k - z^*\|^p$ defined for $p \in [1, \infty)$ (Ortega and Rheinboldt 1970). The sequence $\{z_k\}$ exhibits *linear convergence* if $0 < Q_1 < 1$, *sublinear convergence* if $Q_1 \geqslant 1$, and *superlinear convergence* if $Q_1 = 0$. In this paper, for a more specific characterization of superlinear convergence, we say that the sequence $\{z_k\}$ exhibits *polynomial convergence* if $Q_1 = 0$ and $Q_p > 0$ for some $p > 1$.

We now state the conditions C.1, C.2, and C.3, to be satisfied by the sequences $\{\epsilon_k\}$ and $\{m_k\}$.

C.1. When the numerical procedure used to solve sample-path problems exhibits

(a) linear convergence: $\liminf_{k\to\infty} \epsilon_k \sqrt{m_{k-1}} > 0$;

(b) polynomial convergence: $\liminf_{k\to\infty} (\log(1/\sqrt{m_{k-1}})/\log(\epsilon_k)) > 0$.

C.2. $\limsup_{k\to\infty} (\sum_{j=1}^{k} m_j)\epsilon_k^2 < \infty$.

C.3. $\limsup_{k\to\infty} (\sum_{j=1}^{k} m_j) m_k^{-1} < \infty$.

The conditions C.1, C.2, and C.3 have a clear physical interpretation. The condition C.1 says, for instance, that the sequence $\{\epsilon_k\}$ of error tolerances should not be reduced to zero "too fast" compared to the sequence of sample sizes $\{m_k\}$. Understandably, the notion of "too fast" depends on the convergence rate of the numerical procedure in use, thus warranting the expression of condition C.1 in two parts. The condition C.1(a) is a more stringent special case of C.1(b), i.e., sequences $\{\epsilon_k\}$, $\{m_k\}$ that satisfy condition C.1(a) automatically satisfy C.1(b). This is expected because procedures that exhibit polynomial convergence, as defined, take a smaller far number of "asymptotic" steps to solve to a specific error tolerance than do procedures that exhibit linear convergence. The intuitive sense behind condition C.1 is solving the sample-path problems for only as long as we do not "chase randomness," i.e., only to the extent that it is beneficial from the standpoint of getting closer to the true solution x^*. In order to understand conditions C.2 and C.3, we notice that the number of points visited $N_k \geqslant 1$ for all k. Therefore, during the kth iteration, at least m_k amount of work is done. This work would be wasted if there is no movement in going from the $(k-1)$th retrospective solution X_{k-1} to the kth retrospective solution X_k. This can happen either because ϵ_k is chosen so large that the stipulated tolerance is too easily satisfied, or if the difference between m_{k-1} and m_k is so small that there is little change between the $(k-1)$th sample-path solution X^*_{k-1} and the kth sample-path solution X^*_k. To avoid these scenarios, the conditions C.2 and C.3 impose a lower bound on the rate at which the error-tolerance sequence $\{\epsilon_k\}$ and

the sample-size sequence $\{m_k\}$ go to their respective limits. We discuss specific examples of sequences that satisfy conditions C.1, C.2, and C.3 in §5.3.

As a measure of effectiveness, we use the product of work and squared error. Therefore, at the end of the kth iteration, if the retrospective solution is X_k and the total number of simulation calls expended from iterations 1 through k is W_k, then the random variable of interest is $W_k\|X_k - x^*\|^2$, with smaller values of the random variable being better. See §6 for more on this measure.

We also remind the reader of the stochastic version $O_p(\cdot)$ of the $O(\cdot)$ relation, useful in discussing stochastic convergence. By $X_n = O_p(1)$ we mean that the sequence of random variables $\{X_n\}$ is such that for every $\epsilon > 0$ there exist M_ϵ, N_ϵ such that $F_n(M_\epsilon) - F_n(-M_\epsilon) > 1 - \epsilon$ for all $n > N_\epsilon$, where F_n is the distribution function of X_n. Furthermore, we say X_n is $O_p(Y_n)$ if X_n/Y_n is $O_p(1)$. Therefore, if $\{X_n\}$ and $\{Y_n\}$ are sequences of random variables that converge to zero wp1, and if X_n is $O_p(Y_n)$, we say in loose terms that $\{X_n\}$ converges at least as fast as $\{Y_n\}$.

We are now ready to state two results that establish the sense in which it is beneficial to choose sequences $\{\epsilon_k\}$ and $\{m_k\}$ that satisfy conditions C.1, C.2, and C.3.

THEOREM 5. *Assume that the kth sample-path problem is solved successfully wp*1*, i.e., a retrospective solution X_k is found during the kth iteration so that $\|X_k - X_k^*\| \leqslant \epsilon_k$ wp*1*, where X_k^* is the unique sample-path solution to the kth sample-path problem. Assume that $\{\epsilon_k\} \to 0$ and $\{m_k\} \uparrow \infty$. Let $W_k = \sum_{j=1}^{k} N_j m_j$ be the total number of simulation calls expended from iterations* 1 *through k in solving the sample-path problems, where N_j is the number of points observed during the jth iteration and m_j is the sample-size used to generate the sample-path function during the jth iteration. If the sequences $\{\epsilon_k\}$, $\{m_k\}$ satisfy conditions* C.1, C.2, *and* C.3, *then $W_k\|X_k - x^*\|^2 = O_p(1)$.*

PROOF. In solving the sample-path problem during the kth iteration, the initial solution is X_{k-1}, and the objective is to reach a point within a ball of radius ϵ_k centered on X_k^*. Let N_k be the number of points visited in achieving this. Then, because the sample size used during the kth iteration is m_k, the total work done up to the kth iteration is $W_k = \sum_{j=1}^{k} N_j m_j$. We will now prove the result in two cases characterized by whether the numerical procedure used to solve the sample-path problems exhibits (i) polynomial convergence or (ii) linear convergence.

Case (i) (*polynomial convergence*): Recall that for the kth sample-path problem, $\|X_{k-1} - X_k^*\|$ is the initial distance to the solution, and ϵ_k is the stipulated error tolerance. Because the numerical procedure used for solving the sample-path problem converges superlinearly with Q-order $p > 1$, the number of points visited

$$N_k = O_p\left(1 + \frac{1}{\log p}\left(\log \frac{\log \epsilon_k}{\log \|X_{k-1} - X_k^*\|}\right)\right). \tag{1}$$

The above expression for N_k is obtained upon noting that the errors for a numerical procedure exhibiting polynomial convergence "look like" $\|X_{k-1} - X_k^*\|, \|X_{k-1} - X_k^*\|^p, \|X_{k-1} - X_k^*\|^{p^2}, \ldots$, and then solving for the smallest n such that $\|X_{k-1} - X_k^*\|^{p^n} \leqslant \epsilon_k$.

We know from the CLT on X_k^* (Theorem 4 and Shapiro 2000), and because X_{k-1} is located within a ball of radius ϵ_{k-1} around X_{k-1}^*, that $\|X_{k-1} - X_k^*\| = O_p(1/\sqrt{m_{k-1}} + 1/\sqrt{m_k} + \epsilon_{k-1})$, and so

$$N_k = O_p\left(1 + \frac{1}{\log p}\left(\log \frac{\log \epsilon_k}{\log(1/\sqrt{m_{k-1}} + 1/\sqrt{m_k} + \epsilon_{k-1})}\right)\right).$$

From the above expression for N_k, and because the condition C.1(b) holds,

$$N_k = O_p(1) \quad \text{and} \quad W_k = \sum_{j=1}^{k} N_j m_j = O_p\left(\sum_{j=1}^{k} m_j\right). \tag{2}$$

Again, from the CLT on X_k^* and the relation between X_k and X_k^*, we know that $\|X_k - x^*\|^2 = O_p((1/\sqrt{m_k} + \epsilon_k)^2)$. Therefore, if $\{\epsilon_k\}$ and $\{m_k\}$ are chosen to satisfy conditions C.2 and C.3,

$$W_k\|X_k - x^*\|^2 = O_p\left(\left(\sum_{j=1}^{k} m_j\right)\left(\frac{1}{\sqrt{m_k}} + \epsilon_k\right)^2\right) = O_p(1).$$

Case (ii) (*linear convergence*): Proof is very similar to the previous case, with the condition C.1(a) active instead of C.1(b), and the following alternate expression for N_k instead of (1):

$$N_k = O_p\left(1 + \frac{1}{\log r}\left(\log \frac{\epsilon_k}{\|X_{k-1} - X_k^*\|}\right)\right), \tag{3}$$

where r is a constant satisfying $0 < r < 1$. The expression for N_k is obtained upon noting that the errors for a numerical procedure exhibiting linear convergence "look like" $\|X_{k-1} - X_k^*\|, r\|X_{k-1} - X_k^*\|, r^2\|X_{k-1} - X_k^*\|, \ldots$, and then solving for the smallest n such that $r^n\|X_{k-1} - X_k^*\| \leqslant \epsilon_k$. □

Having shown that conditions C.1, C.2, and C.3 ensure that the random variable of interest $W_k\|X_k - x^*\|^2$ is $O_p(1)$, we now turn to what happens when one or more of the conditions C.1, C.2, and C.3 are violated. We show in Theorem 6 that under Assumption 2, when at least one of the conditions C.1, C.2, or C.3 is violated, the random variable $W_k\|X_k - x^*\|^2 \xrightarrow{p} \infty$.

ASSUMPTION 2. *The sequence*

$$\{(1/\sqrt{m_k} + \epsilon_k)^{-2} \cdot \|X_k - x^*\|^2\} \xrightarrow{d} \Psi,$$

where $\Pr\{\Psi = 0\} = 0$.

In loose terms, Assumption 2 states that the RA algorithm in use does not magically know the location of the

true solution x^* relative to the sample-path solution X_k^*. Because $X_k = X_k^* + \Delta_k$, where Δ_k is a random variable supported on a sphere of radius ϵ_k, the factor $(1/\sqrt{m_k} + \epsilon_k)^{-2}$ is the correct scaling for stabilizing the random variable $\|X_k - x^*\|^2$.

THEOREM 6. *Let the retrospective solutions* $\{X_k\}$ *satisfy Assumption* 2. *Then, if at least one of the conditions* C.1, C.2, *or* C.3 *are violated,* $W_k\|X_k - x^*\|^2 \xrightarrow{p} \infty$.

PROOF. If condition C.1 is violated, and if the limit in the definition of Q-factors exists, the number of points $N_k \xrightarrow{p} \infty$. Because Assumption 2 holds, for given $\nu > 0$ there exists $\delta(\nu) > 0$ so that

$$\Pr\{(1/\sqrt{m_k} + \epsilon_k)^{-2}\|X_k - x^*\|^2 \geqslant \delta(\nu)\} \geqslant 1 - \nu \tag{4}$$

for large enough k. Also,

$$W_k\|X_k - x^*\|^2 \geqslant N_k m_k (1/\sqrt{m_k} + \epsilon_k)^2 \frac{\|X_k - x^*\|^2}{(1/\sqrt{m_k} + \epsilon_k)^2}. \tag{5}$$

Using (4), and because $N_k \xrightarrow{p} \infty$, we see that for any given $M, \nu > 0$,

$$\begin{aligned}
&\Pr\left\{ N_k m_k (1/\sqrt{m_k} + \epsilon_k)^2 \frac{\|X_k - x^*\|^2}{(1/\sqrt{m_k} + \epsilon_k)^2} \leqslant M \right\} \\
&\quad \leqslant \Pr\left\{ \left\{ N_k m_k (1/\sqrt{m_k} + \epsilon_k)^2 \leqslant \frac{M}{\delta(\nu)} \right\} \right. \\
&\qquad\qquad \left. \cup \{(1/\sqrt{m_k} + \epsilon_k)^{-2}\|X_k - x^*\|^2 \leqslant \delta(\nu)\} \right\} \\
&\quad \leqslant \nu + \nu = 2\nu,
\end{aligned} \tag{6}$$

for large enough k, and some $\delta(\nu) > 0$. Conclude from (5) and (6) that $W_k\|X_k - x^*\|^2 \xrightarrow{p} \infty$.

If conditions C.2 or C.3 are violated, the proof is similar—after noticing that

$$W_k\|X_k - x^*\|^2 \geqslant \left(\sum_{j=1}^{k} m_j\right)(1/\sqrt{m_k} + \epsilon_k)^2 \frac{\|X_k - x^*\|^2}{(1/\sqrt{m_k} + \epsilon_k)^2},$$

use $(\sum_{j=1}^{k} m_j)\epsilon_k^2 \to \infty$ when C.2 is violated, or $(\sum_{j=1}^{k} m_j)m_k^{-1} \to \infty$ when C.3 is violated. □

5.2. Convergence of Mean Squared Error

In the previous section, we studied the behavior of the random variable $W_k\|X_k - x^*\|^2$ in assessing the choice of the parameter sequences $\{\epsilon_k\}$ and $\{m_k\}$. We identified three conditions C.1, C.2, and C.3 to be satisfied when choosing the parameter sequences $\{\epsilon_k\}$ and $\{m_k\}$. In this section, in order to further help us identify specific sequences for use, we study parameter choice through the more frequently used mean squared error (MSE). We demonstrate, using Theorems 7 and 8, that it is best to choose sequences $\{\epsilon_k\}$, $\{m_k\}$ that satisfy $0 < \limsup_{k\to\infty} \epsilon_k\sqrt{m_k} < \infty$.

Recall that the mean squared error $\mathrm{MSE}(Z, z^*)$ of a random variable Z with respect to a constant z^* is $\mathrm{MSE}(Z, z^*) = \mathrm{E}[\|Z - z^*\|^2]$. We show through Theorems 7 and 8 that for any given sample-size sequence $\{m_k\}$, to ensure "optimal" convergence of $\mathrm{MSE}(X_k, x^*)$, we should choose the error-tolerance sequence $\{\epsilon_k\}$ so that $\{\epsilon_k^2\}$ converges to zero at the same rate as $\{\mathrm{MSE}(X_k^*, x^*)\}$. We first state Theorem 7, which asserts that it is best to choose the sequence of error tolerances $\{\epsilon_k\}$ so that $\{\epsilon_k^2\}$ converges to zero at least as fast as the sequence $\{\mathrm{MSE}(X_k^*, x^*)\}$.

THEOREM 7. *Assume that the conditions in Theorems* 2 *and* 3 *hold for the SRFP and the SOP contexts, respectively. Also, let the retrospective solutions* $\{X_k\}$ *satisfy* $E[\|X_k - X_k^*\|^2] \geqslant c\epsilon_k^2$ *for some* $c > 0$. *Then*

$$\begin{aligned}
\limsup_{k\to\infty} \frac{\mathrm{MSE}(X_k, x^*)}{\mathrm{MSE}(X_k^*, x^*)} = \infty \quad &\text{if } \limsup_{k\to\infty} \frac{\epsilon_k^2}{\mathrm{MSE}(X_k^*, x^*)} = \infty \\
< \infty \quad &\text{if } \limsup_{k\to\infty} \frac{\epsilon_k^2}{\mathrm{MSE}(X_k^*, x^*)} < \infty.
\end{aligned}$$

PROOF. A proof is provided in the appendix. □

The assumption $\mathrm{E}[\|X_k - X_k^*\|^2] \geqslant c\epsilon_k^2$ for some $c > 0$, which appears in Theorem 7, is rather mild. It implies that because each sample-path problem is solved to within error-tolerance ϵ_k, the deviation of the retrospective solution X_k from the sample-path solution X_k^*, expressed as a fraction of the error-tolerance ϵ_k, remains bounded away from zero. This assumption would be violated only in pathological examples where the retrospective solutions $\{X_k\}$ are such that, across iterations, they are placed within the ϵ_k-ball around X_k^* in progressively closer positions to x^*.

Theorem 7 establishes the *minimum rate* at which the sequence of tolerances $\{\epsilon_k\}$ should converge to zero, i.e., it suggests that the sequence $\{\epsilon_k^2\}$ should converge to zero at least as fast as the sequence $\{\mathrm{MSE}(X_k^*, x^*)\}$. We now present Theorem 8, which is useful in deciding the *maximum rate* of convergence of the sequence of tolerances $\{\epsilon_k\}$. Specifically, Theorem 8 says that as long as the sequence $\{\epsilon_k^2\}$ converges at least as fast as the sequence $\{\mathrm{MSE}(X_k^*, x^*)\}$, irrespective of how much faster it converges, the convergence rate of $\{\mathrm{MSE}(X_k, x^*)\}$ remains the same.

THEOREM 8. *Let the conditions in Theorems* 2 *and* 3 *hold for the SRFP and the SOP contexts, respectively. Let the sequence of error-tolerances* $\{\epsilon_k\}$ *satisfy* $0 < \limsup_{k\to\infty} \epsilon_k^2/(\mathrm{MSE}(X_k^*, x^*))^{1+\nu} < \infty$ *for some* $\nu > 0$. *Then*

$$\limsup_{k\to\infty} \frac{\mathrm{MSE}(X_k, x^*)}{\mathrm{MSE}(X_k^*, x^*)} = 1.$$

PROOF. A proof is provided in the appendix. □

Theorem 8 suggests that there is no gain in the convergence rate of the sequence of mean squared errors

$\{\text{MSE}(X_k, x^*)\}$ with an increase in the convergence rate of the sequence of error tolerances $\{\epsilon_k\}$, as long as $\{\epsilon_k\}$ converges to zero at least as fast as the sequence $\{\text{MSE}(X_k^*, x^*)\}$. This suggests that it may be best to choose the sequence $\{\epsilon_k\}$ so that it tends to zero at the same rate as $\{\text{MSE}(X_k^*, x^*)\}$. This is because choosing $\{\epsilon_k\}$ to converge any faster would mean doing additional work in solving the individual sample-path problems, but without any corresponding benefit in terms of faster convergence of the retrospective solutions $\{X_k\}$. Furthermore, by Theorem 4 for SRFPs, and a corresponding result from Shapiro (2000) for SOPs, we know that $0 < \limsup_{k\to\infty} m_k \text{MSE}(X_k^*, x^*) < \infty$ under mild conditions. This implies that choosing the sequence $\{\epsilon_k\}$ so that it tends to zero at the same rate as $\{\text{MSE}(X_k^*, x^*)\}$ is, in essence, ensuring that $0 < \limsup_{k\to\infty} \epsilon_k\sqrt{m_k} < \infty$.

5.3. A Specific Recommendation for Sample-Size and Error-Tolerance Sequences

In this section, we exploit the results obtained in the two previous sections to recommend specific choices for the sequences $\{\epsilon_k\}$ and $\{m_k\}$. Theorem 9 helps to identify these choices by asserting that the conditions C.1, C.2, and C.3, together with the condition $0 < \limsup_{k\to\infty} \epsilon_k\sqrt{m_k} < \infty$, imply that the sample sizes should exhibit *at most linear growth*, i.e., either sublinear or linear growth, when the numerical procedure in use exhibits linear convergence. Similarly, Theorem 9 also asserts that the sample sizes should exhibit *at most polynomial growth*, i.e., sublinear, linear, or polynomial growth, when the numerical procedure in use exhibits polynomial convergence.

It is worth mentioning that the restriction $0 < \limsup_{k\to\infty} \epsilon_k\sqrt{m_k} < \infty$ obtained in the previous section is redundant when the numerical procedure in use for solving the sample-path problems exhibits linear convergence. This is because it so happens that the conditions C.1(a), C.2, and C.3 automatically imply that $0 < \limsup_{k\to\infty} \epsilon_k\sqrt{m_k} < \infty$. This is, however, not the case when the numerical procedure in use for solving the sample-path problems exhibits polynomial convergence, and the restriction $0 < \limsup_{k\to\infty} \epsilon_k\sqrt{m_k} < \infty$ becomes greatly useful.

THEOREM 9. *Let the sequences* $\{\epsilon_k\}$ *and* $\{m_k\}$ *satisfy* $\{\epsilon_k\} \to 0$, $\{m_k\} \uparrow \infty$, *the conditions* C.1, C.2, C.3, *and* $0 < \limsup_{k\to\infty} \epsilon_k\sqrt{m_k} < \infty$. *Then, if the numerical procedure used to solve the sample-path problems exhibits*

(i) *linear convergence, then* $\limsup_{k\to\infty} m_k/m_{k-1} < \infty$;

(ii) *polynomial convergence, then* $\limsup_{k\to\infty} m_k/m_{k-1}^p < \infty$ *for some* $p \geq 1$.

PROOF. If the numerical procedure used to solve the sample-path problems exhibits linear convergence, the condition C.1(a) is active. Therefore, we have

$$\liminf_{k\to\infty} \frac{\epsilon_k}{1/\sqrt{m_{k-1}}} = \liminf_{k\to\infty} \frac{\epsilon_k\sqrt{m_k}}{\sqrt{m_k/m_{k-1}}} > 0. \tag{7}$$

Table 1. Recommended sample-size growth rates for different convergence rates of the numerical procedure used to solve sample-path problems.

	Exponential growth, e.g., $m_k = \lceil e^{1.1m_{k-1}} \rceil$	Polynomial growth, e.g., $m_k = \lceil m_{k-1}^{1.1} \rceil$	Linear growth, e.g., $m_k = \lceil 1.1m_{k-1} \rceil$
Polynomial conv.	N	Y	Y
Linear conv.	N	N	Y
Sublinear conv.	N	N	NA

Notes. A "Y" indicates that the combination is recommended, an "N" indicates that the combination is not recommended, and an "NA" indicates that a more specific characterization is needed to make a recommendation.

However, because $0 < \limsup_{k\to\infty} \epsilon_k\sqrt{m_k} < \infty$, $\limsup_{k\to\infty} m_k/m_{k-1} = \infty$ would imply that inequality (7) would be violated. Therefore, conclude that assertion (i) holds.

If the numerical procedure used to solve the sample-path problems exhibits polynomial convergence, the condition C.1(b) is active. Therefore, we have

$$\limsup_{k\to\infty} \frac{\log \epsilon_k}{\log(1/\sqrt{m_{k-1}})} < \infty. \tag{8}$$

Because $0 < \limsup_{k\to\infty} \epsilon_k\sqrt{m_k} < \infty$, $\{\epsilon_k\} \to 0$, and $\{m_k\} \to \infty$, there exists $C > 0$ such that $\epsilon_k < C/\sqrt{m_k}$ and $|\log(\epsilon_k)| > |\log(C/\sqrt{m_k})|$ for large enough k. Using this result with (8), we note that $\limsup_{k\to\infty} \log(C/\sqrt{m_k})/\log(1/\sqrt{m_{k-1}}) < \infty$. This implies that $\limsup_{k\to\infty} \log(m_k)/\log(m_{k-1}) < \infty$. Conclude, after noting $m_k \geqslant m_{k-1}$ for all k, that $\limsup_{k\to\infty} m_k/m_{k-1}^p < \infty$ for some $p \geqslant 1$. □

A natural choice for the sequence of sample sizes exhibiting linear growth is $m_0 = 1$, $m_k = \lceil cm_{k-1} \rceil$, where c is some chosen constant greater than one. This ensures that sample size is increased by $100(c-1)\%$ during each iteration. Similarly, a choice for the sequence of sample sizes exhibiting polynomial growth is $m_0 = 1$, $m_k = \lceil m_{k-1}^c \rceil$, where c is some constant greater than one. For the sequence of error tolerances $\{\epsilon_k\}$, a natural choice resulting from the stipulation $0 < \limsup_{k\to\infty} \epsilon_k\sqrt{m_k} < \infty$ is $\epsilon_k = K/\sqrt{m_k}$, where K is some constant greater than zero. It is easy to verify that the combination $m_0 = 1$, $m_k = \lceil cm_{k-1} \rceil$, $\epsilon_k = K/\sqrt{m_k}$ satisfies conditions C.1(a), C.2, and C.3; whereas the combination $m_0 = 1$, $m_k = \lceil m_{k-1}^c \rceil$, $\epsilon_k = K/\sqrt{m_k}$, satisfies conditions C.1(b), C.2, and C.3. Table 1 further clarifies the recommended choices of $\{\epsilon_k\}$, $\{m_k\}$.

6. Postscript

In this section we discuss two issues surrounding our analysis of sample-path algorithms. First, recall that we used $W_k\|X_k - x^*\|^2$ for measuring algorithm efficiency. Why has this form been chosen? Why not use $W_k\|X_k - x^*\|$ or $W_k\|X_k - x^*\|^3$, or a more elaborate measure such as

the "expected computational work done until we attain a solution X_k that is $\delta > 0$ within x^* with 95% confidence"? Second, we have assumed that the individual iterations within an RA algorithm are terminated by checking if $\|X_k - x^*\| \leqslant \epsilon_k$. What if this check cannot be performed directly?

In what follows, we discuss both of these issues. We argue somewhat subjectively in §6.1 that our choice of efficiency measure has broad appeal from the standpoint of intuitiveness, asymptotics, and ease of analysis. In §6.2 we take up the issue of detecting termination and provide heuristic strategies when a direct check of the termination criterion is unavailable.

6.1. A Justification for the Chosen Measure of Efficiency

Recall from §5.2 that the sequence of retrospective solutions $\{X_k\}$ cannot converge to x^* any faster than the sequence of sample-path solutions $\{X_k^*\}$. We also know from Theorem 4 (and corresponding results for SOPs) that, under general conditions, $\{X_k^*\}$ converges to x^* at the canonical rate $O(1/\sqrt{m_k})$. We thus see that unless sample paths are generated in an intelligent non-i.i.d. fashion, the fastest-possible convergence rate for the sequence $\{X_k\}$ is $O(1/\sqrt{m_k})$.

This last point motivates the measure $W_k\|X_k - x^*\|^2$. Because $\|X_k - x^*\|^2$ cannot converge faster than $O(1/m_k)$, and the work done up to the kth iteration W_k is at least m_k, the best we can achieve through intelligent parameter choice is ensuring that $W_k\|X_k - x^*\|^2$ does not tend to infinity. By contrast, the measure $W_k\|X_k - x^*\|$ tends to infinity *irrespective of parameter choice*. Thus, in considering the form $W_k\|X_k - x^*\|^r$, $r = 2$ is the smallest exponent that remains interesting by providing opportunities for efficiency gains through parameter choice. Values of r larger than two are less discerning because if $W_k\|X_k - x^*\|^2$ is asymptotically finite, so is $W_k\|X_k - x^*\|^r$ for $r > 2$.

Instead of measures having the form $W_k\|X_k - x^*\|^r$, we could have used a measure such as "the expected work done to ensure with 95% confidence that the retrospective solution X_k is at most δ within x^*." Such a measure is undoubtedly appealing, and has been pursued recently (Bayraksan and Morton 2010) in the context of identifying *feasible* sampling plans that guarantee identifying solutions of a probabilistic standard. However, performing optimization across parameters using such a finite-time measure poses analytical challenges due to difficulties associated with characterizing the expected work done at termination. Specifically, whereas making statements about guaranteed stopping is possible (as shown in Bayraksan and Morton 2010), deriving expressions for the exact number of iterations until stopping is difficult.

In all of the above-discussed measures, what constitutes W_k? As detailed in §5.1, W_k is the total number of simulation calls expended in visiting the various points up to the kth iteration. For instance, let $k = 3$, $m_1 = 1$, $m_2 = 2$, $m_3 = 4$, and suppose the algorithm visited three points in the first iteration, two in the second, and two in the third. Then $W_k = (3 \times 1) + (2 \times 2) + (2 \times 4) = 15$. Implicit in such a calculation are two key points: (i) the computational work involved in performing algebraic operations is negligible compared to the computational work due to simulation calls; (ii) simulation calls are not differentiated by function, e.g., we have made no attempt to specially deal with possibly less burdensome simulation calls made expressly for gradient or Hessian calculations. Although recognizing both of these as drawbacks, we note that the main results of the paper remain unchanged if these sources of error have at most a linear effect on the total work done, i.e., have the effect of inflating (or deflating) the total work done by only a proportionality constant.

6.2. Detecting Termination

Recall the way error has been characterized—as the distance $\|X_k - X_k^*\|$ in the solution space. This choice is convenient because it allows easy exposition in terms of the convergence rates (invariably expressed in the solution-space; see, for example, Ortega and Rheinboldt 1970, Chapter 9) of the deterministic numerical procedure that is used to solve the sample-path problems. Alternatively, we could have constructed our results by measuring error in the function space—as $\|\bar{y}_{m_k}(X_k; \underline{\omega}_k) - \bar{y}_{m_k}(X_k^*; \underline{\omega}_k)\| = \|\bar{y}_{m_k}(X_k; \underline{\omega}_k)\|$ in the SRFP context, and as $\|\nabla\bar{y}_{m_k}(X_k; \underline{\omega}_k) - \nabla\bar{y}_{m_k}(X_k^*; \underline{\omega}_k)\| = \|\nabla\bar{y}_{m_k}(X_k; \underline{\omega}_k)\|$ in the SOP context. Results analogous to Theorems 5, 6, 7, and 8 then follow directly, and the exact same conclusions on the relation between the choice of parameters and convergence rates hold true, but with convergence rates now interpreted in the function space as opposed to the solution space.

From an implementation standpoint, when the condition $\|X_k - X_k^*\| \leqslant \epsilon_k$ cannot be verified directly, we recommend a heuristic check based on a first-order approximation of the function $\bar{y}_{m_k}(x; \underline{\omega}_k)$. Specifically, for SRFPs, we have

$$\begin{aligned} 0 &= \bar{y}_{m_k}(X_k^*; \underline{\omega}_k) \\ &= \bar{y}_{m_k}(X_k; \underline{\omega}_k) + H(X_k)(X_k^* - X_k) + R(X_k^* - X_k), \end{aligned} \tag{9}$$

where $H(X_k)$ is the gradient of $\bar{y}_{m_k}(x; \underline{\omega}_k)$ at X_k, and $R(\cdot)$ is a reminder term (Ortega and Rheinboldt 1970, p. 184) satisfying $\lim_{h\to 0} R(h)/\|h\| = 0$. Using (9), terminate the kth iteration in SRFPs if $\|\widehat{H}(X_k)^{-1}\bar{y}_{m_k}(X_k; \underline{\omega}_k)\| \leqslant \epsilon_k$, where $\widehat{H}(X_k)$ is an estimate of $H(X_k)$. Similarly, in the SOP context, terminate the kth iteration if $\|\widehat{H}(X_k)^{-1}\widehat{\nabla}\bar{y}_{m_k}(X_k; \underline{\omega}_k)\| \leqslant \epsilon_k$, where $\widehat{H}(X_k)$ is an estimate of the Hessian $H(X_k)$ of $\bar{y}_{m_k}(x; \underline{\omega}_k)$ at X_k, and $\widehat{\nabla}\bar{y}_{m_k}(X_k; \underline{\omega}_k)$ is an estimate of the gradient $\nabla\bar{y}_{m_k}(X_k; \underline{\omega}_k)$ of $\bar{y}_{m_k}(x; \underline{\omega}_k)$ at X_k. The essential idea is that, assuming that the feasible set $\mathfrak{D}$ contains all local minima in its interior, and assuming that the Hessian $H(X_k^*)$ is positive definite, $X_k - X_k^*$ can be approximated by $\widehat{H}(X_k)^{-1}\widehat{\nabla}\bar{y}_{m_k}(X_k; \underline{\omega}_k)$ when X_k is sufficiently close to X_k^*.

Admittedly, this heuristic checks only the first-order conditions and has no knowledge about whether the second-order sufficient conditions for a minimum are satisfied at X_k^*. It thus breaks down, for example, when the Hessian $H(X_k^*)$ is singular or when X_k is close to a local maximum. For more elaborate termination heuristics, including strategies for checking second-order sufficient conditions in smooth problems and general optimality conditions in nonsmooth problems, see Gill et al. (1986, §8.2.3).

7. A Numerical Example

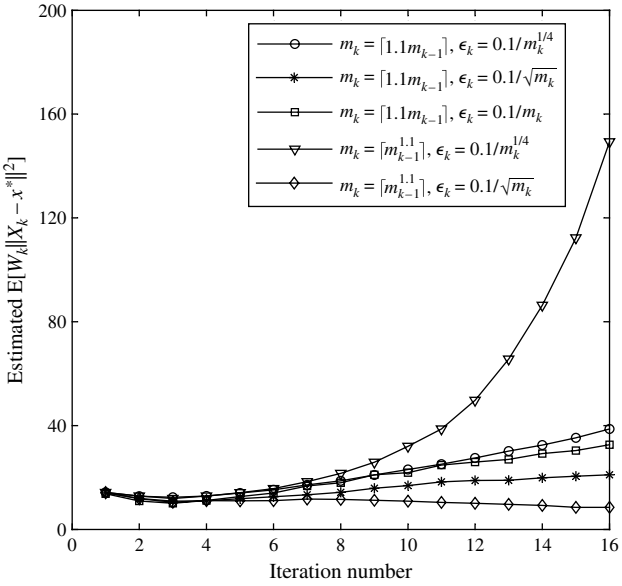
We now illustrate the main results of this paper through a stochastic version of the analytic center (AC) problem (Boyd 2004, §8.5.3). Recall that the AC problem involves finding the point with the maximum depth in a bounded region formed by a set of linear constraints. Formally, the AC problem involves maximizing $\prod_{i=1}^{M} S_i$, subject to $Ax \leqslant B$, $|x_i| \leqslant h$, $x \in \mathbb{R}^q$, where A is an $M \times q$ matrix, B is an $M \times 1$ vector of positive constants, h is a positive constant, and x_i is the ith element of x. Also, $S_i = B_i - A_i x$, where B_i is the ith element of the vector B, and A_i is the ith row of matrix A.

The stochastic analytic center (SAC) problem is a stochastic analogue of the AC problem obtained by replacing the number of constraints M, and the elements in A and B, with random variables, and then maximizing $\mathrm{E}[\prod_{i=1}^{M} S_i]$.

For the specific numerical example considered in this section, the number of constraints M is generated from a distribution that takes on values $1, 2, \ldots, q$ with equal probability. The elements in the vector B are generated independently, each from an exponential distribution with mean 1. The (i, j)th element of the matrix A is formed as $A(i, j) = U(i, j)/\sqrt{\sum_{j=1}^{q} U^2(i, j)}$, where $U(i, j)$ is uniformly distributed between $j - 1$ and j. The sample-path problems are solved using the guarded Newton method with backtracking search (Boyd 2004, pp. 464, 487).

Recall that Theorems 5, 6, and 9 show that when the sequences $\{m_k\}$ and $\{\epsilon_k\}$ are chosen to satisfy conditions C.1, C.2, and C.3, the random variable $W_k\|X_k - x^*\|^2$ is $O_p(1)$. Figure 1 illustrates this by plotting an estimate of $\mathrm{E}[W_k\|X_k - x^*\|^2]$ as a function of the iteration number k for various choices of $\{m_k\}$ and $\{\epsilon_k\}$. Because the numerical procedure used to solve the sample-path problems in this example exhibits polynomial convergence, the results in this paper recommend that sample sizes grow *at most polynomially*, and $\epsilon_k = K/\sqrt{m_k}$. Among the five $\{m_k\}$, $\{\epsilon_k\}$ choices for which curves are plotted in Figure 1, only $m_0 = 1$, $m_k = \lceil 1.1 m_{k-1} \rceil$, $\epsilon_k = 0.1/\sqrt{m_k}$, and $m_0 = 1$, $m_k = \lceil m_{k-1}^{1.1} \rceil$, $\epsilon_k = 0.1/\sqrt{m_k}$, satisfy the recommendations. Accordingly, we see that the recommended choices dominate the others. (In Figure 1, instead of plotting the random variable $W_k\|X_k - x^*\|^2$, we plot an estimate of its expectation. It can be shown that if $W_k\|X_k - x^*\|^2 \xrightarrow{d} \Psi$, the sequence $\{W_k\|X_k - x^*\|^2\}$ is uniformly integrable, and $\mathrm{E}[\Psi] < \infty$, then $\lim_{k\to\infty} \mathrm{E}[W_k\|X_k - x^*\|^2] = \mathrm{E}[\Psi]$. Also, if $W_k\|X_k - x^*\|^2 \xrightarrow{p} \infty$, then $\lim_{k\to\infty} \mathrm{E}[W_k\|X_k - x^*\|^2] = \infty$.)

Figure 1. For the SAC example, following the guidelines in this paper, the error tolerance should be the inverse square-root of the sample size, and sample sizes can grow at most polynomially.



Note. Only the second and the fifth of the five parameter choices displayed satisfy these conditions, and they dominate the rest.

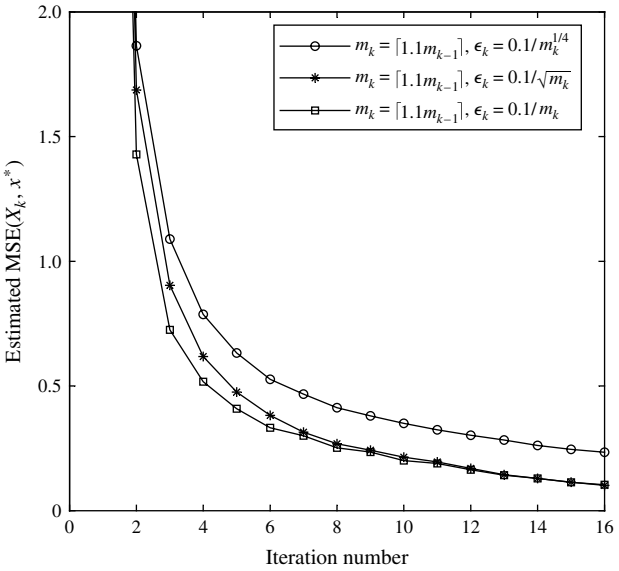
Figure 2 is a plot of the estimated mean squared error $\mathrm{MSE}(X_k, x^*)$ as a function of the iteration number k. It explicitly omits the number of simulation calls, and illustrates Theorems 7 and 8. Recall that Theorems 7 and 8 show that for any given sample-size sequence $\{m_k\}$, choosing $\{\epsilon_k\}$ to converge to zero at the same rate as the sequence $\{\mathrm{MSE}(X_k^*, x^*)\}$ maximizes the convergence rate of the sequence $\{\mathrm{MSE}(X_k, x^*)\}$. This implies, specifically, that choosing $\{\epsilon_k\}$ to converge to zero any faster than $\{1/\sqrt{m_k}\}$ is of no benefit. This is apparent from the plots in Figure 2.

8. Summary and Concluding Remarks

RA methods for solving SRFPs and SOPs generate a sequence of sample-path problems with increasing sample sizes $\{m_k\} \to \infty$, and then solve them to decreasing error tolerances $\{\epsilon_k\} \to 0$. Under mild conditions, detailed in this paper, RA iterates converge to the correct solution wp1.

A more challenging issue that this paper addresses is providing guidance in choosing the error-tolerance sequence $\{\epsilon_k\}$, and the sample-sizes sequence $\{m_k\}$, used by RA algorithms. We argue, with theoretical support, that these sequences should be chosen to satisfy $0 < \limsup_{k\to\infty} \epsilon_k \sqrt{m_k} < \infty$, and the three conditions C.1, C.2, and C.3 characterized in this paper. In broad terms, these stipulations imply that the rate at which sample sizes are

Figure 2. In RA algorithms, choosing the error-tolerance sequence $\{\epsilon_k\}$ to reduce to zero at the same rate as $\{1/\sqrt{m_k}\}$ maximizes the convergence rate of the sequence of mean squared errors $\{\mathrm{MSE}(X_k, x^*)\}$.



Note. Choosing $\{\epsilon_k\}$ to converge any faster provides no additional benefit.

increased across iterations in RA algorithms should depend on the speed of convergence of the numerical procedure used to solve the sample-path problems. Specifically, when the numerical procedure exhibits linear convergence, sample sizes should grow at most linearly, and when the numerical procedure exhibits polynomial convergence, the sample sizes should grow at most polynomially.

Two additional remarks relating to future research are now in order.

(i) The guidelines for parameter choice in this paper, although very useful, still leave unchosen constants for the implementer. We suspect that the quality of these constants will depend on the specific instance of the SRFP or SOP on hand. Accordingly, methods that dynamically adjust sample-size and error-tolerance choices as the algorithm evolves, and while staying within the prescribed convergence rates, will be very beneficial from an efficiency standpoint.

(ii) The theory presented in this paper assumes i.i.d. generation of sample paths. In practice, although i.i.d. sample-path generation is easiest and (therefore) widely followed, "intelligent" correlated generation of sample paths can achieve great efficiency gains. The theory corresponding to such convergence is only now appearing for SOPs (Homem-de-Mello 2008) and is virtually nonexistent for SRFPs.

Appendix

Proof of Theorem 2. Let $h(x) = \|g(x)\|^2$, and its estimator $\bar{w}_{m_k}(x; \underline{\omega}_k) = \|\bar{y}_{m_k}(x; \underline{\omega}_k)\|^2$. The function $h(x)$ is continuous, and the sequence $\{\bar{w}_{m_k}(x; \underline{\omega}_k)\} \to h(x)$ uniformly wp1. Furthermore, the set π^* of zeros of $g(x)$ coincides with the set of minimizers of $h(x)$, and X_k^* is a zero of $\bar{y}_{m_k}(x; \underline{\omega}_k)$ if and only if it is also a global minimizer of $\bar{w}_{m_k}(x; \underline{\omega}_k)$. Therefore, applying Theorem 1 to function $h(x)$ and its estimator $\bar{w}_{m_k}(x; \underline{\omega}_k)$, we see that $\mathrm{dist}(X_k^*, \pi^*) \to 0$ wp1. However, because $\mathrm{dist}(X_k, \pi^*) \leqslant \mathrm{dist}(X_k^*, \pi^*) + \|X_k - X_k^*\| \leqslant \mathrm{dist}(X_k^*, \pi^*) + \epsilon_k$, and $\epsilon_k \to 0$, we conclude $\mathrm{dist}(X_k, \pi^*) \to 0$ wp1. □

Proof of Theorem 4. By Taylor's Theorem (Rudin 1976, p. 110), we know that $\bar{y}_{m_k}(X_k^*; \underline{\omega}_k) = \bar{y}_{m_k}(x^*; \underline{\omega}_k) + \nabla \bar{y}_{m_k}(\xi_k; \underline{\omega}_k)(X_k^* - x^*)$, where ξ_k lies on the line joining X_k^* and x^*. Because $\bar{y}_{m_k}(X_k^*; \underline{\omega}_k) = 0$, this implies $\sqrt{m_k}(X_k^* - x^*) = -\nabla \bar{y}_{m_k}(\xi_k; \underline{\omega}_k)^{-1}(\sqrt{m_k}\bar{y}_{m_k}(x^*; \underline{\omega}_k))$ for large enough k wp1. Also, because $X_k^* \to x^*$ wp1, $\nabla \bar{y}_{m_k}(x; \underline{\omega}_k) \to \nabla g(x)$ uniformly wp1 (in some neighborhood of x^*), and $\nabla g(x^*)$ is nonsingular, we know that $\nabla \bar{y}_{m_k}(X_k^*; \underline{\omega}_k)^{-1} \to \nabla g(x^*)^{-1}$ wp1. Combining these with the CLT on $\bar{y}_{m_k}(x; \underline{\omega}_k)$, we conclude that Theorem 4 holds. □

Proof of Theorem 7. Recall that X_k is the kth retrospective solution, X_k^* is the kth sample-path solution, and $\|X_k - X_k^*\| \leqslant \epsilon_k$ wp1. Then, noting that $\mathrm{MSE}(X_k, x^*) = \mathrm{E}[\|X_k - x^*\|^2]$, we write

$$\mathrm{MSE}(X_k, x^*) = \mathrm{E}[\|X_k - X_k^*\|^2] + \mathrm{MSE}(X_k^*, x^*) + 2|\mathrm{E}[(X_k - X_k^*)^T (X_k^* - x^*)]|. \tag{10}$$

Defining the inner product $\langle X, Y \rangle = \mathrm{E}[X^T Y]$ (Szechtman 2006, p. 263), and using the Cauchy-Schwarz inequality $|\langle X, Y \rangle| \leqslant \sqrt{\langle X, X \rangle}\sqrt{\langle Y, Y \rangle}$ (Ortega and Rheinboldt 1970, p. 39) on the last term in expression (10), we get

$$|\mathrm{E}[(X_k - X_k^*)^T (X_k^* - x^*)]| \leqslant \sqrt{\mathrm{E}[\|X_k - X_k^*\|^2]}\sqrt{\mathrm{MSE}(X_k^*, x^*)}. \tag{11}$$

Using the inequality (11) and the expression (10), we get

$$\left| \frac{\mathrm{MSE}(X_k, x^*)}{\mathrm{MSE}(X_k^*, x^*)} - \frac{\mathrm{E}[\|X_k - X_k^*\|^2]}{\mathrm{MSE}(X_k^*, x^*)} - 1 \right| \leqslant 2\sqrt{\frac{\mathrm{E}[\|X_k - X_k^*\|^2]}{\mathrm{MSE}(X_k^*, x^*)}}. \tag{12}$$

Also, because $\mathrm{E}[\|X_k - X_k^*\|^2] = c_k \epsilon_k^2$ with $1 \geqslant c_k \geqslant c > 0$,

$$\frac{c\epsilon_k^2}{\mathrm{MSE}(X_k^*, x^*)} \leqslant \frac{\mathrm{E}[\|X_k - X_k^*\|^2]}{\mathrm{MSE}(X_k^*, x^*)} \leqslant \frac{\epsilon_k^2}{\mathrm{MSE}(X_k^*, x^*)}. \tag{13}$$

Conclude from inequalities (12) and (13) that the assertion in Theorem 7 holds. □

Proof of Theorem 8. Because $0 < \limsup_{k\to\infty} \epsilon_k^2/(\mathrm{MSE}(X_k^*, x^*))^{1+\nu} < \infty$ for some $\nu > 0$, we know that

$\limsup_{k\to\infty} \epsilon_k^2/\mathrm{MSE}(X_k^*, x^*) = 0$. Using this in inequality (13) appearing in the proof of Theorem 7, we get $\limsup_{k\to\infty} \mathrm{E}[\|X_k - X_k^*\|^2]/\mathrm{MSE}(X_k^*, x^*) = 0$. Using this in inequality (12) appearing in the proof of Theorem 7, we conclude that the assertion in Theorem 8 holds. □

References

Andradóttir, S. 2006. An overview of simulation optimization via random search. S. G. Henderson, B. L. Nelson, eds. *Simulation. Handbooks in Operations Research and Management Science*. Elsevier, Amsterdam, 617–631.

Atlason, J., M. A. Epelman, S. G. Henderson. 2004. Call center staffing with simulation and cutting plane methods. *Ann. Oper. Res.* **127** 333–358.

Atlason, J., M. A. Epelman, S. G. Henderson. 2008. Optimizing call center staffing using simulation and analytic center cutting-plane methods. *Management Sci.* **54**(2) 295–309.

Barton, R. R., M. Meckesheimer. 2006. Metamodel-based simulation optimization. S. G. Henderson, B. L. Nelson, eds. *Simulation. Handbooks in Operations Research and Management Science*. Elsevier, Amsterdam, 535–574.

Bastin, F., C. Cirillo, P. L. Toint. 2006. Convergence theory for nonconvex stochastic programming with an application to mixed logit. *Math. Programming* **108** 207–234.

Bayraksan, G., D. P. Morton. 2007. Assessing solution quality in stochastic programs. *Math. Programming Ser. B* **108** 495–514.

Bayraksan, G., D. P. Morton. 2010. A sequential sampling procedure for stochastic programming. *Oper. Res.* Forthcoming.

Bhatnagar, S., V. S. Borkar. 1998. A two time scale stochastic approximation scheme for simulation based parametric optimization. *Probab. Engrg. Informational Sci.* **12** 519–531.

Bhatnagar, S., M. C. Fu, S. I. Marcus, S. Bhatnagar. 2001. Two-timescale algorithms for simulation optimization of hidden Markov models. *IIE Trans.* **33** 245–258.

Boyd, S. 2004. *Convex Optimization*. Cambridge University Press, Cambridge, UK.

Chen, H., B. W. Schmeiser. 2001. Stochastic root finding via retrospective approximation. *IIE Trans.* **33** 259–275.

Fu, M. C. 2002. Optimization for simulation: Theory vs. practice. *INFORMS J. Comput.* **14** 192–215.

Gill, P. E., W. Murray, M. H. Wright. 1986. *Practical Optimization*. Elsevier, London.

Glasserman, P. 1991. *Gradient Estimation via Perturbation Analysis*. Kluwer, Dordrecht, The Netherlands.

Healy, K., L. W. Schruben. 1991. Retrospective simulation response optimization. B. L. Nelson, D. W. Kelton, G. M. Clark, eds. *Proc. 1991 Winter Simulation Conf.*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, 954–957.

Herer, Y. T., M. Tzur, E. Yucesan. 2006. The multilocation transshipment problem. *IIE Trans.* **38** 185–200.

Higle, J., S. Sen. 1991. Stochastic decomposition: An algorithm for two-stage linear programs with recourse. *Math. Oper. Res.* **16**(3) 650–669.

Homem-de-Mello, T. 2003. Variable-sample methods for stochastic optimization. *ACM Trans. Modeling Comput. Simulation* **13** 108–133.

Homem-de-Mello, T. 2008. On rates of convergence for stochastic optimization problems under non-i.i.d. sampling. *SIAM J. Optim.* **19**(2) 524–551.

Homem-de-Mello, T., A. Shapiro, M. L. Spearman. 1999. Finding optimal material release times using simulation-based optimization. *Management Sci.* **45**(1) 86–102.

Kim, S.-H., B. L. Nelson. 2006. Selecting the best system. S. G. Henderson, B. L. Nelson, eds. *Simulation. Handbooks in Operations Research and Management Science*. Elsevier, Amsterdam, 501–534.

Kleywegt, A. J., A. Shapiro, T. Homem-de-Mello. 2001. The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* **12** 479–502.

Kushner, H., D. Clark. 1978. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York.

Kushner, H. J., G. G. Yin. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag, New York.

Mak, W. K., D. P. Morton, R. K. Wood. 1999. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Oper. Res. Lett.* **24** 47–56.

Nemirovski, A., A. Shapiro. 2004. Convex approximations of chance constrained programs. *Optim. Online*. http://www.optimization-online.org/.

Ólafsson, S. 2006. Metaheuristics. S. G. Henderson, B. L. Nelson, eds. *Simulation. Handbooks in Operations Research and Management Science*. Elsevier, Amsterdam, 633–654.

Ortega, J. M., W. C. Rheinboldt. 1970. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York.

Pasupathy, R. 2006. On choosing parameters in retrospective-approximation algorithms for simulation-optimization. L. Perrone, F. Wieland, J. Liu, B. Lawson, D. Nicol, R. Fujimoto, eds. *Proc. 2006 Winter Simulation Conf.*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, 208–215.

Pasupathy, R., B. W. Schmeiser. 2009. Retrospective-approximation algorithms for multidimensional stochastic root-finding problems. *ACM TOMACS* **19**(2) 5:1–5:36.

Plambeck, E. L., B. R. Fu, S. M. Robinson, R. Suri. 1996. Sample-path optimization of convex stochastic performance functions. *Math. Programming* **75** 137–176.

Polak, E., J. O. Royset. 2008. Efficient sample sizes in stochastic nonlinear programming. *J. Comput. Appl. Math.* **217**(2) 301–310.

Prakash, P., G. Deng, M. C. Converse, J. G. Webster, G. M. Mahvi, M. C. Ferris. 2008. Design optimization of a robust sleeve antenna for hepatic microwave ablation. *Phys. Med. Biol.* **53** 1057–1069.

Robbins, H., S. Monro. 1951. A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407.

Robinson, S. 1996. Analysis of sample-path optimization. *Math. Oper. Res.* **21** 513–528.

Rubinstein, R. Y., A. Shapiro. 1993. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley & Sons, New York.

Rudin, W. 1976. *Principles of Mathematical Analysis*. Mc-Graw Hill, New York.

Shapiro, A. 1991. Asymptotic analysis of stochastic programs. *Ann. Oper. Res.* **30** 169–186.

Shapiro, A. 2000. Stochastic programming by Monte Carlo simulation methods. *Stochastic Programming E-Print Ser.* http://hera.rz.hu-berlin.de/speps/.

Shapiro, A. 2004. Monte Carlo sampling methods. A. Ruszczynski, A. Shapiro, eds. *Stochastic Programming. Handbooks in Operations Research and Management Science*. Elsevier, Amsterdam, 353–426.

Spall, J. C. 2003. *Introduction to Stochastic Search and Optimization*. John Wiley & Sons, Inc., Hoboken, NJ.

Spall, J. C. 2006. Feedback and weighting mechanisms for improving Jacobian (Hessian) estimates in the adaptive simultaneous perturbation algorithm. *Proc. Amer. Control Conf.*, Piscataway, NJ, June.

Szechtman, R. 2006. A Hilbert space approach to variance reduction. S. G. Henderson, B. L. Nelson, eds. *Simulation. Handbooks in Operations Research and Management Science*. Elsevier, Amsterdam, 259–290.

Tadic, V., S. P. Meyn. 2003. Asymptotic properties of two time-scale stochastic approximation algorithms with constant step sizes. *Proc. Amer. Control Conf.*, Piscataway, NJ, June.

Verweij, B., S. Ahmed, A. Kleywegt, G. Nemhauser, A. Shapiro. 2003. The sample average approximation method applied to stochastic vehicle routing problems: A computational study. *Comput. Appl. Optim.* **24** 289–333.

Wardi, Y., B. Melamed, C. G. Cassandras, C. G. Panayiotou. 2002. Online IPA gradient estimators in stochastic continuous fluid models. *J. Optim. Theory Appl.* **115**(2) 369–405.